**NTNU – Trondheim**
Norwegian University of
Science and Technology

# Is (Deep) Neural Networks Trustworthy?

Hao Wang, Big Data Lab, IIR, NTNU, Norway

Panel for ISPA 2017, Guangzhou, China, Dec 14, 2017

# On my background

- Ph.D. and B.Eng. in Computer Engineering
- 6 years in Canada
  - 4 years in Nova Scotia (St. Francis Xavier Univ.)
    - Workflow modeling and HPC
  - 2 years in IBM Canada/McMaster Univ.
    - Software Certification, Safety Analysis
- Moved to Norway in 2014
  - Big Data Lab
    - Applied R&D and Education on Big Data (Visual Analytics) and IIoT
    - Applications from Ocean (Maritime and Aquaculture), Sustainability (Water and Air pollution), Health (Cancer), FinTech

# NTNU Ålesund

- Closely integrated with local industry
- Strong interdisciplinary collaborative atmosphere

Barents Sea
Hammerfest
Murmansk
Tromsø
Norwegian Sea
Kiruna
White Sea
Arctic Circle
Luleå
Oulu
NORWAY
Umeå
FINLAND
Trondheim
SWEDEN
Gulf of Bothnia
Tampere
Lake Ladoga
Turku
Helsinki
Gävle
Gulf of Finland
St. Petersburg
Bergen
Oslo
Stockholm
ÅLAND ISLANDS
Tallinn
ESTONIA
Stavanger
Göteborg
Gotland
Riga
LATVIA
North Sea
DENMARK
Copenhagen
Malmö
Öland
Baltic Sea
LITHUANIA
Vilnius
Vitsyebsk
Mahilyow
Minsk
Kaliningrad
RUSSIA
BELARUS
Bornholm
Gdańsk
Hrodna
Hamburg
Warsaw
Brest

# Sunnmøre Maritime Cluster

**A world-leading maritime cluster**

- 13 design companies

  14 ship yards

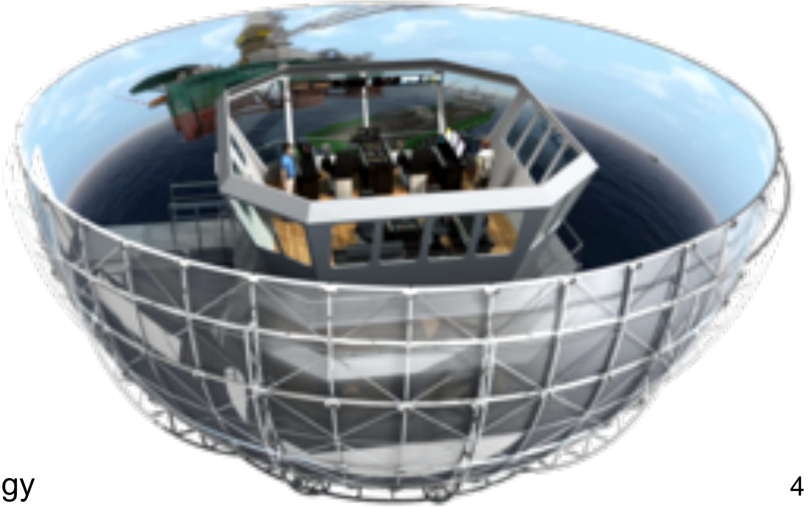
  20 ship-owner companies

  169 equipment suppliers

  22 500 employees

  40% of the world's modern fleet

*Most of Norway's strength in the field of advanced marine operations is concentrated within an hour's drive from Ålesund. The region is home for an impressive constellation of over 200 leading maritime companies and training, research and finance institutions that form one of the most complete maritime clusters in the world.*



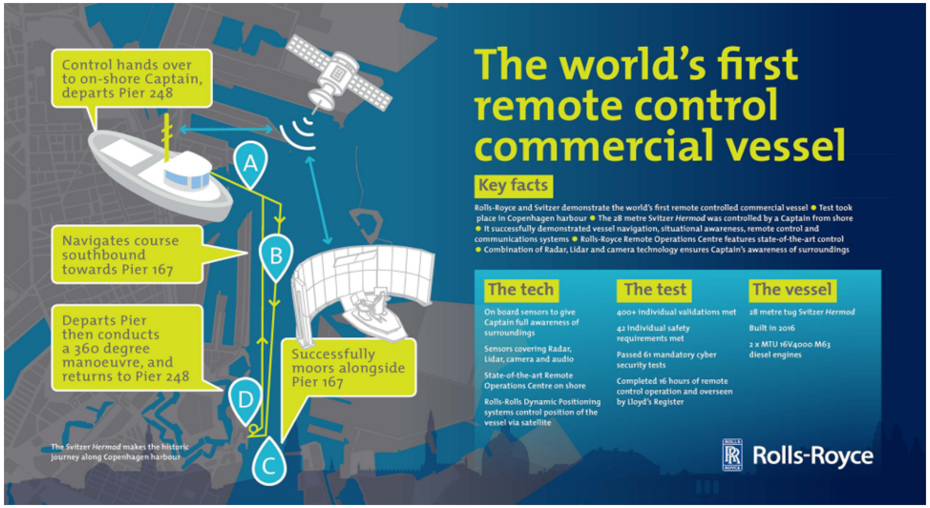
# Data Analytics for Maritime Operations

- Sensor data obtained through *HEalth MOnitoring System* (HEMOS) by Rolls Royce Marine AS
  - Time series + GPS
  - High frequency
    - E.g., vibration, torque
  - Low frequency
    - E.g., heading, speed, and GPS

# Security for Cyber-enabled Vessels

- A project with Profs. Sokratis K. Katsikas, Slobodan Petrovic, and Edmund Førland Brekke
- Goals
  - To analyze the cyber security risks of the cyber-enabled remotely-operated ship
  - To propose effective and efficient risk management strategies

June 2017: Rolls-Royce and Svitzer demonstrate world's first remotely operated commercial vessel in Copenhagen habour


Control hands over to on-shore Captain, departs Pier 248
A
Navigates course southbound towards Pier 167
B
Departs Pier then conducts a 360 degree manoeuvre, and returns to Pier 248
D
Successfully moors alongside Pier 167
C
The Svitzer *Hermod* makes the historic journey along Copenhagen harbour
Rolls-Royce

The world's first remote control commercial vessel

Key facts

Rolls-Royce and Svitzer demonstrate the world's first remote controlled commercial vessel ● Test took place in Copenhagen harbour ● The 28 metre Svitzer *Hermod* was controlled by a Captain from shore ● It successfully demonstrated vessel navigation, situational awareness, remote control and communications systems ● Rolls-Royce Remote Operations Centre features state-of-the-art control ● Combination of Radar, Lidar and camera technology ensures Captain's awareness of surroundings

The tech

On board sensors to give Captain full awareness of surroundings

Sensors covering Radar, Lidar, camera and audio

State-of-the-art Remote Operations Centre on shore

Rolls-Rolls Dynamic Positioning systems control position of the vessel via satellite

The test

400+ individual validations met

42 individual safety requirements met

Passed 61 mandatory cyber security tests

Completed 16 hours of remote control operation and overseen by Lloyd's Register

The vessel

28 metre tug Svitzer *Hermod*

Built in 2016

2 x MTU 16V4000 M63 diesel engines

# Robustness of Deep Neural Networks

- NN are being integrated in Safety Critical Systems
- Szegedy et al. [1] observed that state- of-the-art NN are highly vulnerable to *adversarial perturbations*
  - Given a correctly-classified input x, it is possible to find a new input x′ that is very similar to x but is as- signed a different label.
  - For instance, in image-recognition networks it is possible to add a small amount of noise (undetectable by the human eye) to an image and change how it is classified by the network [2].

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks, 2013. Technical Report.

[2] Divya Gopinath, Guy Katz, Corina S. Pasareanu, and Clark Barrett. DeepSafe: A Data-driven Approach for Checking Adversarial Robustness in Neural Networks. 2017

# Hypotheses on this Weakness

- 1. high complexity and non-linearity of neural networks can assign random labels in areas of the space which are under-explored [1].
  - Has been refuted
    - Unable justify the transferability of adversarial samples from one model to another
    - Linear models also suffer from this phenomenon

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks, 2013. Technical Report.

# Hypotheses on this Weakness

- 2. Goodfellow et al. [3] proposed a *linearity hypothesis* instead:
  - deep neural networks are highly non-linear with respect to their parameters, but mostly linear with respect to their inputs, and adversarial examples are easy to encounter when exploring a direction orthogonal to a decision boundary.
  - Another conjecture to explain the existence of adversarial examples is the cumulation of errors while propagating the perturbations from layer to layer.
    - A small carefully crafted perturbation in the input layer may result in a much greater difference in the output layer, effect that is only magnified in high dimensional spaces, causing the activation of the wrong units in the upper layers

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2014.

# Approaches

- Casting to solve an optimization problem

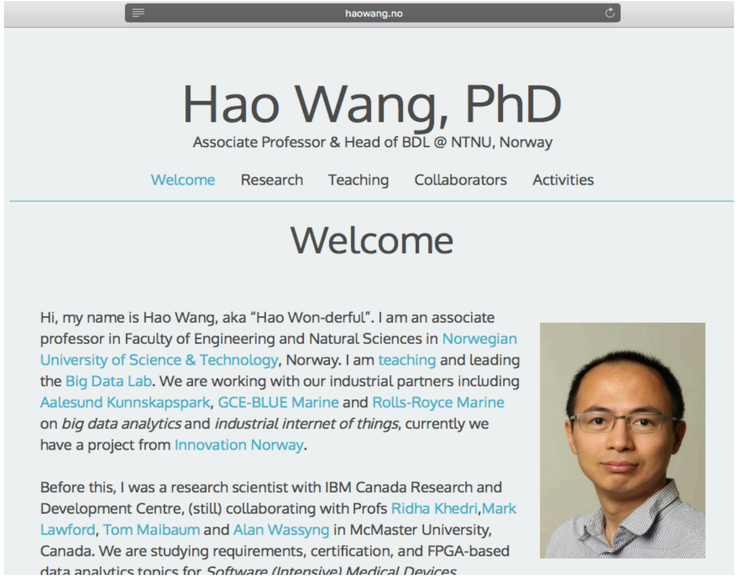
$$\min_{\Delta x} F(x + \Delta x) \neq F(x), \quad \text{s.t.} \quad x' \in [0, 1]^n.$$

- Verification approach
  - E.g., Deep Learning Verification (DLV) is an approach that defines a region of safety around a known input and applies SMT solving for checking robustness

[4] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu. Safety verification of deep neural networks. In Proc. 29th Int. Conf. on Computer Aided Verification (CAV), pages 3–29, 2017

# Tusen takk!!

- Comments?
- Collaborative ideas?



## Hao Wang, PhD

Associate Professor & Head of BDL @ NTNU, Norway

## Welcome

Hi, my name is Hao Wang, aka "Hao Won-derful". I am an associate professor in Faculty of Engineering and Natural Sciences in Norwegian University of Science & Technology, Norway. I am teaching and leading the Big Data Lab. We are working with our industrial partners including Aalesund Kunnskapspark, GCE-BLUE Marine and Rolls-Royce Marine on *big data analytics* and *industrial internet of things*, currently we have a project from Innovation Norway.

Before this, I was a research scientist with IBM Canada Research and Development Centre, (still) collaborating with Profs Ridha Khedri,Mark Lawford, Tom Maibaum and Alan Wassyng in McMaster University, Canada. We are studying requirements, certification, and FPGA-based data analytics topics for *Software (Intensive) Medical Devices*

- Hao Wang, hawa@ntnu.no, www.haowang.no
- Big Data Lab http://blog.hials.no/bigdata/